手机短消息过滤系统的研究与实现

林 青　史晓东

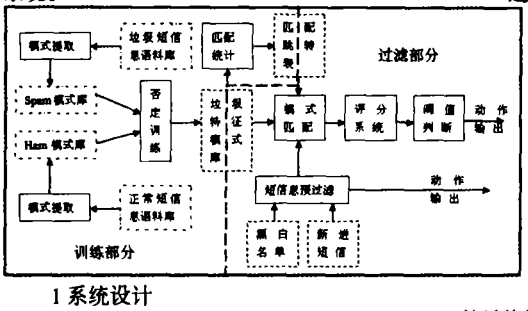
（厦门大学计算科学系，福建 厦门 361005）

摘　要：如何处理手机短信息垃圾已经成为日益困扰广大用户的一个难题。介绍了一个基于智能手机平台的短信息过滤系统，基于模式分析-模式匹配的过滤方案。采用模式分析的方法自动对垃圾短信息语料库进行模式提取。并给出了系统的整体测试情况及改进方向。

关键词：手机短消息；过滤系统；研究；实现

引言。手机已经成为大多数人不可缺少的通讯工具。手机短信已经成为人们进行沟通和交流的新手段。同时，短信息的火爆使其安全性受到了前所未有的挑战。用户在享受短信给自己带来的方便的同时，也受到各种恶意信息的干扰。因此进行垃圾短信息过滤的研究显得尤为有必要。

近年来以 Smartphone 和 Symbian 系统为代表的智能手机终端开始广泛流行，为直接在手机终端上自动处理短信息提供了必要的硬件平台；文本分类技术的日益成熟，为研究提供了理论支撑。因此我们基于 Smartphone 手机平台，使用文本分类技术来构建手机短消息过滤系统。



1 系统设计

系统采用了基于模式提取-模式匹配的方法。先收集一定数量的垃圾短信息输入系统进行学习训练，从中提取出垃圾短信息所包含的特征模式。而后在收到新短信息时，将垃圾特征模式与信息文本进行比较，从而对文本是否垃圾信息作出判断。

1.1 学习训练部分

该部分工作于 PC 终端上。本部分主要由模式提取、否定训练、匹配统计三个功能模块组成。

1.1.1 模式提取模块

模式提取的作用是从输入的语料中提取出模式信息。工作原理基于 TEIRESIAS 算法[1]。这是一种模式分析算法，可以用于在大量的信息数据中快速发现未知模式。

1.1.2 否定训练模块

从 Spam 集中提取出来的模式中既有垃圾短信息中特有的模式特征，也含有正常短信息中也会出现的模式特征。必须去除这些非垃圾特有的模式特征以免出现虚假判断。

去除的方法是将正常的短信息集同样交由模式提取模块，提取出正常短信息的模式特征，若些特征在 Spam 库中存在，则将其从 Spam 库中移除。

1.1.3 匹配统计模块

本模块生成匹配跳转表，用于辅助在模式匹配阶段的工作，实现快速的多模式多点匹配。

1.2 信息过滤部分

信息过滤部分运行于手机终端上，在收到新的短信息的时候，由系统提交给信息过滤模块检查。模块在检查之后做出判断，若判断为正常短信息则给予放行；若判断为垃圾信息则直接删除。

该部分主要由信息预处理，模式匹配，评分系统三个主要模块组成。

1.2.1 信息预处理

短信息预过滤的任务是预过滤新收取到的短信息，对于可以通过短信息头直接判断的直接进行处理。对于无法使用短信息头判断的则提交后继模块进行内容判断。

1.2.2 模式匹配

这是过滤模块的核心部分。其功能是对信息正文进行多模式匹配，找出信息中所有存在的与垃圾模式库中相匹配的模式。

我们借鉴 Aho-Corasick 算法思想，改进了字符等价关系的定义，使之适合于多模式多点匹配。从而实现了多模式多点的线性时间复杂度匹配。

1.2.3 评分系统

由评分系统对模式匹配结果进行打分，将其分数结果将与阈值作比较，并最终输出判定结果。

我们分别计算两个值，一是匹配出来的模式数量/文本长度，记为 S1；二是匹配出来的所有模式中所有非任选字符数量之和/文本长度，记为 S2。

2 测试结果及分析

我们使用 Ling-Spam 作为测试语料库，在测试中我们综合观察训练模块的训练参数 L，W 和过滤模块的 S1，S2，共 4 个参数的影响。

反垃圾系统不同于其它文本分类系统，它可以容忍较低的召回率，而对于将有用信息错误过滤掉则被认为是不可接受的。所以我们在测试中只考虑 Precission > 99%情况下的 Recall。

部分测试结果，按 Recall 的高低排序如下：

W	L	S1	S2	Precission(%)	Recall(%)
4	3	0.023256	0.093023	99.0279	71.3415
6	3	0.02439	0.097561	99.0022	66.0081
7	4	0.017857	0.071429	99.0124	64.909

可以看到，采用(W, L) = (4, 3)的训练参数时，可以达到最高的召回率为 71.3%。

3 结语

对在手机客户端进行短信息过滤的进行了一定的探索。提出了一个短信息过滤系统的模型，并对其实际过滤效果进行了测试。结果是比较令人满意的。

但是，目前关于这方面的研究刚刚起步，只能借鉴文本分类、反垃圾邮件等方面的技术进行研究，还有很多可改进之处，例如：

3.1 可以尝试引入权重信息来进一步提高识别率。

3.2 可以尝试引入特征选择来缩小模式库大小，并且进一步提高过滤效率和识别率。

另外，垃圾短信息语料库的缺乏也对研究造成了一定的障碍，这些都有待未来的进一步研究改进。

参考文献

[1]Chung-Kwei: a Pattern-discovery-based System for theAutomatic Identification of Unsolicited E-mail Messages.

[2]Digital cellular telecommunications system (Phase 2+);Use of Data Terminal Equipment - Data Circuit terminating; Equipment (DTE - DCE) interface for Short Message Service (SMS) and Cell Broadcast Service (CBS)(GSM 07.05 version 5.5.0) .

[3]A. V. Aho , M. J. Corasick. Efficient String Matching: An Aid to Bibliographic Search [J]. Communications of the ACM, 1975,18(6):333-340.

（上接 30 页）

3.1 开发能与 5GM168 集成芯片所实现的功能相同的芯片电路

5GM168 集成芯片实现的功能是将放大后的音频信号整流再放大，通过控制压控振荡器的振荡频率，利用脉冲分配器输出 0 和 1 的数字信号。

在本课题的后续研究开发中可以用数字和模拟电路的知识设计车此电路，可提供给厂商集成化成为新开发的芯片资源。

3.2 可编程的多方式色彩控制自动电路

本课题的后续研究的另一个方向可以将人工和自动相结合的方式实行排练编程，演出时自动控制的可编程多方式色彩自动控制研究。

这有些和软件控制灯光效果相接轨，但是以后的电子电路的研究都是往软硬件结合的方向发展，所以本课题的这个后续研究方向是可行的。

参考文献

[1]夏路易，石宗义.电路原理图与电路板设计教程 Protel99SE[M].北京：希望电子出版社，2002.

[2]门宏.精选电子制作 66 例[M].人民邮电出版社.

[3]彭介华.电子技术课程设计指导[M].北京：高等教育出版社，1997.

[4]童诗白，华成英.模拟电子技术基础[M].北京：高等教育出版社，2001.

作者简介：陈刚（1983,10,4~），现就读于乐山师范学院，本科学历，研究方向为应用电子。

责任编辑：李光旭

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net